

# An empirical analysis of the Ebola outbreak in West Africa

Abdul Khaleque<sup>1,\*</sup> and Parongama Sen<sup>1,†</sup>

<sup>1</sup>*Department of Physics, University of Calcutta, 92 APC Road, Kolkata 700009, India*

The data for the Ebola outbreak that occurred in 2014-2016 in three countries of West Africa are analysed within a common framework. The analysis is made using the results of an agent based Susceptible-Infected-Removed (SIR) model on a Euclidean network, where nodes at a distance  $l$  are connected with probability  $P(l) \propto l^{-\delta}$  in addition to nearest neighbors. The cumulative density of infected population here has the form  $R(t) = \frac{a \exp(t/T)}{1 + c \exp(t/T)}$ , where the parameters depend on  $\delta$  and the infection probability  $q$ . This form is seen to fit well with the data. Using the best fitting parameters, the time at which the peak is reached is estimated and is shown to be consistent with the data. We also show that in the Euclidean model, one can choose  $\delta$  and  $q$  values which reproduce the data for the three countries qualitatively. These choices are correlated with population density, control schemes and other factors.

PACS numbers: 87.19.Xx, 07.05.Kf, 95.75.-z

## I. INTRODUCTION

Mathematical modelling of the phenomena of disease spreading has a long history, the first such attempts being made in the early twentieth century. Typically, an individual is assumed to be in either in one of the three possible states: susceptible, infected and removed (or recovered) denoted by S, I, and R respectively in the simplest models. Diseases which can be contracted only once are believed to be described by the SIR model in which a susceptible individual gets infected by an infected agent who is subsequently removed (dead or recovered). A removed person no longer takes part in the dynamics. In SIS model, an infected person may become susceptible again. Plenty of variations and modifications of the SIR and SIS models have been considered over the last few decades. Resurgence of interest in these models has taken place following the discovery that social networks do not behave like random or regular networks [1, 2]. The recent emphasis has been to study these models on complex networks like small world and scale free networks. A few surprising results have been derived theoretically in the recent past [2].

In mathematical models, one quantifies the infection probability. In most theoretical models, the epidemic has a threshold behaviour as the infection probability is varied. However, an estimate of this quantity from real data is difficult as it is related to biological features like nature of the pathogen etc.

The test of a model lies in its ability to match real data. Not appreciable success has been made so far although some qualitative consistency has been achieved [2, 3]. The available data is usually in the form of number of newly infected patients and total (cumulative) number of cases. In the SIR model, the newly infected fraction shows an initial growth followed by a peak and a sub-

sequent decay. This matches with the overall structure of the real data (e.g. for Severe Acute Respiratory Syndrome (SARS)), which however, show local oscillatory behaviour in addition. Such a behaviour may be due to demographic non uniformity [4].

It is meaningful to study the epidemic spreading by considering that the agents are embedded on an Euclidean space. A few models on Euclidean networks have been studied earlier which show that the geographical factor plays an important role in the spreading process [5–13]. In particular, the SIR model on an Euclidean network where the agents may be connected to a few randomly chosen long range neighbours with a probability decaying with the Euclidean distance has been considered in some detail [12, 13].

In 2014, the Ebola virus caused large scale outbreaks mainly in three West African countries and only recently it has been declared as over (June 2016). Ebola virus is transmitted through body fluids and blood and it is also believed that a person can contract the disease only once. A few attempts have been made to analyse the data so far [14–19]. Different factors like demographic effect, hospitalization, vaccination and treatment plans have been incorporated in the traditional and well-known SIR model to understand the dynamics of Ebola disease [17–19]. However, in these models, a mean field approximation has been used which is rather unphysical. Using the results of an agent based SIR model on Euclidean network mentioned in the last paragraph [12], we have analysed the Ebola data for the three countries Guinea, Liberia and Sierra Leone in West Africa where the outbreak extended over approximately two years.

In section II, we discuss the details of the available data and the method of the analysis. Analysis of the data and simulation results are presented in section III and in the last section summary and discussions are made.

\*Email: aktphys@gmail.com

†Email: parongama@gmail.com

TABLE I: Statistics of Ebola data for three different Countries.

Country	Total Cases	Density of Infected Population	Lab-Confirmed Cases	Total Deaths
Guinea	3814	$3.0 \times 10^{-4}$	3358	2544
Sierra Leone	14124	$2.2 \times 10^{-3}$	8706	3956
Liberia	10678	$2.2 \times 10^{-3}$	3163	4810
Total	28616		15227	11310

## II. METHOD

We consulted the Ebola data for the number of cases detected in the three countries Guinea, Liberia and Sierra Leone in West Africa (The Centers for Disease Control and Prevention (CDC) [20]). The data is available from 25th March 2014 to 13th April 2016 at the time interval of a few days. The data is noisy and contains obvious errors as often the cumulative data is shown to decrease. The first available data is from March 2014 when Guinea was already struck with the disease for some time (first case in Guinea reported in December 2013) such that the data for the initial period is missing. For Liberia and Sierra Leone, the data for initial stage are available, however these are sparse and unreliable; often the data for number of death exceeds the number of cases. For this reason, the data has been analysed from the date when the number of cases detected is at least 50 for each country. Even then the errors cannot be fully avoided as for very late stages, the data being rare, also become somewhat unreliable. Hence, the entire data set has to be handled carefully.

In Table I, a summary of the statistics of the Ebola data is presented and one can immediately note that all cases could not have been confirmed in the laboratory in the case of Liberia where number of deaths exceeds the laboratory confirmed cases. Obviously many cases were unreported. For Guinea, these two figures are closest and the data for Guinea is in fact the cleanest one. We have studied the available data for total (cumulative) number of cases  $R(t)$  as a function time  $t$  and extracted the data for number of new cases  $I(t)$  from these.

Another point needs to be mentioned. The disease has been officially declared over on 1st June 2016 for Guinea, 9th June 2016 for Liberia and 17th March 2016 for Sierra Leone [21]. But one can see from Fig. 1 that the cumulative data shows a saturation over fairly long period of time. Apparently a few stray cases delayed the declaration of the disease being over. For Liberia, for example, the disease was originally declared to be over as early as in May 2015 but two small flare-ups were reported later. However the cumulative data is hardly affected by the later cases.

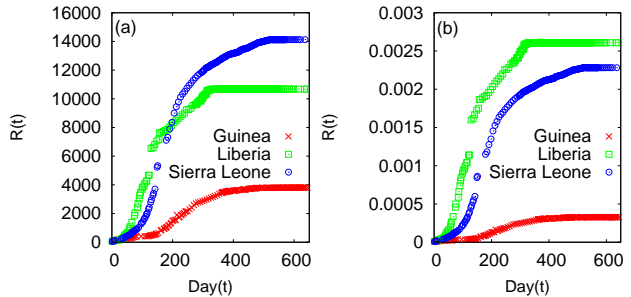


FIG. 1: (a) Cumulative number of infected individuals as a function of time (day) for the three countries Guinea, Liberia and Sierra Leone. (b) Same data normalised by the population of each country.

## III. RESULTS

### A. Data analysis

Most of the data analyzed available so far deal with actual numbers of cases. However, as we attempt to provide a comparative picture, we have taken the fraction i.e., divided the numbers by the total population for each country. One can easily see that the comparative trends become different in the two different approaches (Fig. 1). The disease is seen to affect the least fraction of the population in Guinea and the maximum in Liberia. However, the number of cases is maximum for Sierra Leone. On the other hand, the disease has existed over a longer period in the case of Sierra Leone and Guinea, considering the time at which the data reach a saturation value.

Had the infection probability been the sole factor responsible for the spread, the patterns would have been the same for the three countries. Hence we argue that to compare the data to a theoretical model the latter must have more than one parameter. A minimal model would consist of two parameters like the one considered in [12]. Here the agents have two nearest neighbour connections and a random long range connection to an agent located at a distance  $l$  with probability  $l^{-\delta}$ . The other parameter is of course the infection probability  $q$ . This study revealed that above a threshold value of  $q(\delta)$  an epidemic can occur.

The removed population in the model was fitted to the form:

$$R(t) = \frac{a \exp(t/T)}{1 + c \exp(t/T)}, \quad (1)$$

where  $a$ ,  $c$  and  $T$  depend on the values of  $\delta$  and  $q$ . Note that the removed population in the model essentially corresponds to the cumulative infected cases since in the model the infected agents were assumed to be removed immediately after being infected. Hence this fitting form is used for the cumulative data of infected cases and shows very good agreement for Guinea (Fig. 2), while

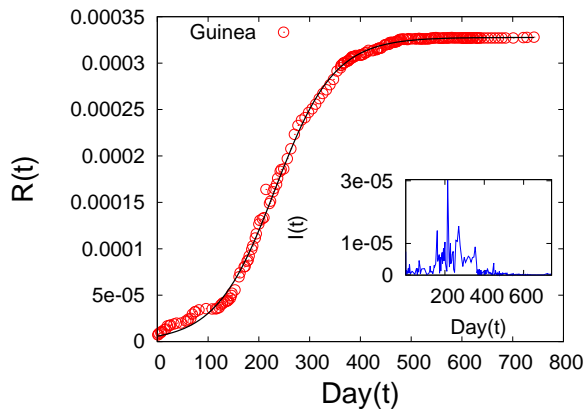


FIG. 2: Cumulative fraction of population infected and the fitted curve as a function of time (day) for country Guinea. Inset is for the fraction of newly infected population as a function of time.

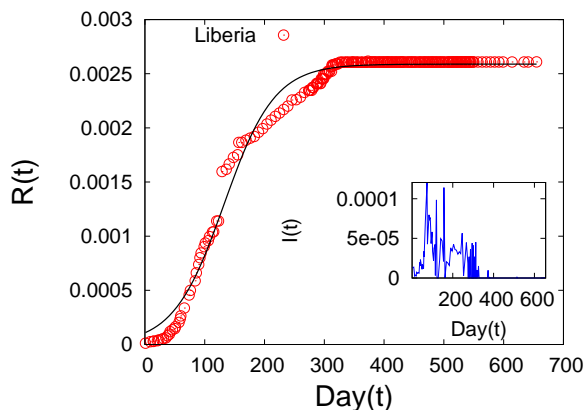


FIG. 3: Cumulative fraction of population infected and the fitted curve as a function of time (day) for country Liberia. Inset is for the fraction of newly infected population as a function of time.

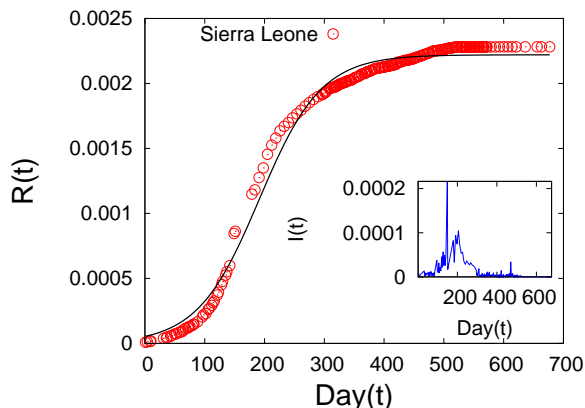


FIG. 4: Cumulative fraction of population infected and the fitted curve as a function of time (day) for country Sierra Leone. Inset is for the fraction of newly infected population as a function of time.

TABLE II: Exponents  $a$ ,  $c$  and  $T$  for three different Countries.

Country	$a$	$c$	$T$	$t_p$
Guinea	0.0000059	0.0182146	57.915	231.98
Liberia	0.0001125	0.0434763	42.1957	132.30
Sierra Leone	0.0000549	0.0247653	51.758	191.41

there is fairly good agreement with the data of the other two countries (Figs. 3,4). From eq.1, one can show that a peak value for  $I(t)$  will occur at  $t_p = T \log(1/c)$  [12]. The associated values of the exponents  $a$ ,  $c$  and  $T$  are found out for the three countries and the values of  $t_p$  also extracted. The latter is to be compared to the peak occurring in the (newly) infected fraction against time plotted in the insets of the Figures 2,3 and 4. These data shows a lot of fluctuation and not a very clear peak but the theoretically estimated values of  $t_p$  tally with a large value of new cases occurring close to this time. The exponent values and  $t_p$  are tabulated in Table II. The errors in the estimation of exponents are  $\mathcal{O}(10^{-6})$  for  $a$ ,  $\mathcal{O}(10^{-3})$  for  $c$  and  $\mathcal{O}(10^0)$  for  $T$ . One can see that  $t_p$  is also directly proportional to the total duration, being least for Liberia and maximum for Guinea.

## B. Results from the model

The cumulative data for infected people has a sigmoid form in general and has been shown to have a form given by eq (1) in a recent study as well [19]. To establish that indeed the Euclidean network is a proper model responsible for the epidemic spreading, one should be able to reproduce from the model the consistent results and trends using appropriate values of the parameters, at least qualitatively.

The Euclidean model has been already described in Sec. III A and was first considered in [12]. Initially all nodes are susceptible and one arbitrary node is chosen and taken to be infected. Time is discretized and infected people can infect susceptible individuals with probability  $q$  and will be removed within one unit of time with the assumption that they are either dead or cured. In the present simulation, for the single network, 400 such choices have been taken and quantities are averaged. A secondary averaging is made by considering 100 different network configurations. Periodic boundary condition has been used in the simulation. Systems with size  $N = 2^{11}$  has been considered.

The behaviour of the network depends on the value of  $\delta$ . The network behaves as a small world network for  $\delta < 1$  and as a regular one dimensional lattice for  $\delta > 2$ . For  $1 < \delta < 2$ , it shows short range behaviour. One should use a value of  $\delta$  larger than 1 as the Ebola virus spreads through actual body contact such that the underlying network must be short ranged. Also  $\delta < 2$  is

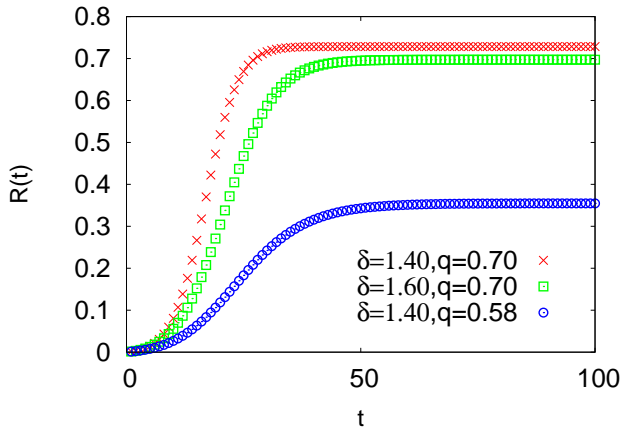


FIG. 5: Fraction of population infected as a function of time (Monte Carlo time step) for different pairs of infection rate  $q$  and  $\delta$ .

chosen as a real network is more connected than a regular one. The values of  $q$  should be same in principle. However, the value of  $q$  may be effectively altered using control schemes like contact tracing, quarantining the patient and efficiently treating the disease. Such possibilities have not been directly included in the model. We will address this issue in the last section again.

We first discuss the case of Liberia and Sierra Leone. We note that the saturation values are quite close while the saturation in Liberia has been reached earlier. Here we find that the same value of  $q$  but a different value of  $\delta$  can indeed reproduce these features; Fig. 5 shows the results for  $\delta = 1.4$  for Liberia and 1.6 for Sierra Leone while the  $q$  values are same ( $q = 0.70$ ).

We next discuss the case for Guinea. It has the lowest saturation value of the cumulative data for infected population while the disease is of duration slightly larger than that of Sierra Leone. This makes it quite apparent that one has to use a smaller value of  $q$  to get data consistent with that of Guinea. We find that indeed one can get such values of  $q$  keeping  $\delta = 1.4$  such that the saturation value is smaller while the duration is larger comparatively. We show the data in Figs. 5 and 6. The values of  $t_p$  are shown in Table III are also consistent with the real data. The argument behind the choices of  $\delta$  and  $q$  are discussed in the next section.

#### IV. DISCUSSIONS

We have analysed the data for Ebola outbreak in West African countries which are available for 2014-16 and also reproduced qualitatively the data using a model of epidemic spreading. In this section we justify the choice of the parameters used in the model to obtain the results

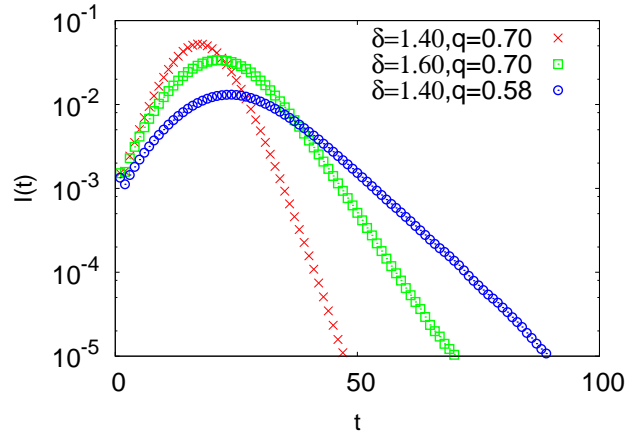


FIG. 6: Fraction of newly infected population as a function of time (Monte Carlo time step) for different pairs of infection rate  $q$  and  $\delta$ .

TABLE III: Exponents  $a, c$  and  $T$  for different values of parameters

Parameters (Country)	$a$	$c$	$T$	$t_p$
$\delta = 1.4, q = 0.70$ (Liberia)	0.005116	0.006974	3.39171	16.8415
$\delta = 1.6, q = 0.70$ (Sierra Leone)	0.011130	0.015718	5.10908	21.2175
$\delta = 1.4, q = 0.58$ (Guinea)	0.011842	0.032342	7.09658	24.3510

consistent with the real data. We have already justified the choice of  $\delta$  between 1 and 2 in the last section. We have used a larger value of  $\delta$  for Sierra Leone and a smaller value of  $q$  for Guinea to get the consistency.

To justify why  $\delta$  should be larger for Sierra Leone we note the following. Sierra Leone and Liberia are comparable in size but the density of population is much higher in the former. The density of population is  $79.4/\text{km}^2$  and  $40.43/\text{km}^2$  respectively for the two countries [22]. Hence the number of neighbours within the same distance is larger for Sierra Leone which implies a larger value of  $\delta$  effectively.

On the other hand, the population densities of Guinea ( $40.90/\text{km}^2$ ) [23] and Liberia are quite close so that one should use the same  $\delta$  value. However, we need to justify why a smaller value of  $q$  is able to reproduce the data for Guinea. A smaller value of  $q$  indicates less infection probability which is possible if proper medical care and control measurements are taken. This is indeed true as we find from several documents that the disease was tackled most effectively in Guinea. Table I clearly shows that the maximum percentage of cases for Guinea were laboratory-tested which indicates that the process of contact tracing and treatment were more efficient. This is supported by the fact that in Guinea, about 56 con-

tacts per infected person were traced compared to 23 in case of Sierra Leone [24]. We find from [25] that MSF treated the largest number of reported cases in Guinea, in Sierra Leone the minimum out of reported cases. Thus most cases in Sierra Leone, even when reported, had received less attention while in Liberia, a large number is not confirmed or reported at all. Apparently, medical centers by international organisations have also been set up much earlier in Guinea as it was the epicenter of the disease and the disease started as early as in 2013 December. However, later activities could control the disease in Liberia and Sierra Leone as well, and the final number of deaths had been far less than initially anticipated. We also note a curious fact - though Guinea may have recorded the minimum number of cases, yet the disease spanned a longer duration compared to Liberia. Further analysis, beyond the scope of the present paper, may be able to explain this.

We add here a few more relevant comments. We note that while qualitative features of the data obtained from the model are quite similar to the real data, quantitatively they are much larger. This may be due to the fact that for the real data, the entire population has been

taken to obtain the fractions while the disease might not have prevailed in such totality due to geographical or other factors. We have also made simple assumptions like homogeneity, i.e., uniform number of contacts for all agents. The initial condition has been taken to be identical: the disease commences with only one infected people. Our assumption that agents are immobile is supported by [19] in which it is argued that migration does not play a role in the spreading. Even so, this simple model is able to yield data which is consistent with real data.

The effect of the ebola outbreak has been devastating in the West African countries. Apart from the human losses, economic loss has also been considerable [26]. The present study shows that the Euclidean model can be treated as a basic starting point and can be further developed by adding other features. This will make it very useful and important for making accurate predictions. Work is in progress towards that direction.

Acknowledgement: PS acknowledges support from CSIR grant.

- 
- [1] A. Barrat, M. Barthelemy and A. Vespignani, *Dynamical processes on complex networks* (Cambridge University Press, Cambridge, U.K., 2008).
  - [2] P. Sen and B. K. Chakrabarti, *Sociophysics: An Introduction*, Oxford University Press, 2013.
  - [3] H. W. Hethcote, SIAM Review **42**, 599 (2000).
  - [4] H. W. Hethcote, Mathematical Problems in Biology, Lecture Notes in Biomathematics, Springer, Berlin **2**, 83 (1974).
  - [5] H. K. Janssen and K. Oerding and F. Van Wijland and H. J. Hilhorst, The European Physical Journal B **7**, 137 (1999).
  - [6] F. Linder, J. Tran-Gia, S. R. Dahmen and H. Hinrichsen, Journal of Physics A **41**, 185005 (2008).
  - [7] S. N. Bennett, A. J. Drummond, D. D. Kapan, M. A. Suchard, J. L. Munoz-Jordan, O. G. Pybus, E. C. Holmes and D. J. Gubler, Molecular biology and evolution **27**, 811 (2010).
  - [8] Z. Wu, K. Rou and H. Cui, AIDS Education and Prevention **16**, 7 (2004).
  - [9] X. Xu, H. Peng, X. Wang and Y. Wang, Physica A **367**, 525 (2006).
  - [10] J. Wang, Z. Liu and J. Xu, Physica A **382**, 715 (2007).
  - [11] Z. Zhao, Y. Liu and M. Tang, Chaos **22**, 023150 (2012).
  - [12] A. Khaleque and P. Sen, J. Phys. A: Math. Theor. **46**, 095007 (2013).
  - [13] P. Grassberger, Journal of Statistical Mechanics: Theory and Experiment **2013**, P04004 (2013).
  - [14] J. A. Lewnard, M. L. N. Mbah, J. A. Alfaro-Murillo, F. L. Altice, L. Bawo, T. G. Nyenswah and A. P. Galvani, The Lancet Infectious Diseases **14**, 1189 (2014).
  - [15] D. Chowell, C. Castillo-Chavez, S. Krishna, X. Qiu and K. S. Anderson, The Lancet Infectious Diseases **15**, 148 (2015).
  - [16] A. Camacho, A. J. Kucharski, S. Funk, P. Piot and W. J. Edmunds, Epidemics **9**, 70 (2014).
  - [17] A. Rachah and D. F. M. Torres, Discrete Dynamics in Nature and Society **2015**, 842792 (2015).
  - [18] A. Radulescu, J. Herron, arXiv:1512.06305, (2015).
  - [19] K. Burghardt, C. Verzijl, J. Huang, M. Ingram, B. Song and M. Hasne arXiv:1606.07497, 842792 (2016).
  - [20] <https://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/previous-case-counts.html>.
  - [21] [https://en.wikipedia.org/wiki/West\\_African\\_Ebola\\_virus\\_epidemic](https://en.wikipedia.org/wiki/West_African_Ebola_virus_epidemic).
  - [22] <https://en.wikipedia.org/wiki/Liberia>, [https://en.wikipedia.org/wiki/Sierra\\_Leone](https://en.wikipedia.org/wiki/Sierra_Leone).
  - [23] <https://en.wikipedia.org/wiki/Guinea>.
  - [24] <http://ebolaresponse.un.org/guinea>, <http://ebolaresponse.un.org/sierra-leone>.
  - [25] [https://www.doctorswithoutborders.org/sites/usa/files/msf\\_ebola\\_accountability\\_report\\_final\\_05\\_11\\_2016\\_2\\_002.pdf](https://www.doctorswithoutborders.org/sites/usa/files/msf_ebola_accountability_report_final_05_11_2016_2_002.pdf).
  - [26] <https://www.brookings.edu/wp-content/uploads/2016/07/fighting-ebola-songwe-2.pdf>.